Linear Regression

Dr. Sarah Hunter

3/25/2020

Bivariate Regression

The workhorse of data analysis is linear regression. Linear regression is used as a hypothesis test when you have a continuous dependent variable. In some cases, you want to know the relationship between two variables, the dependent variable of which is a continuous variable. In this case, you can use bivariate (meaning two variables) regression. R's base functions can perform linear regression quite easily.

Recall that linear regression takes the data and fits a linear model by estimating the regression parameters. In other words, linear regression takes the input (the X variable(s)) and the output (the Y variable) and estimates the following equation:

$$Y_i = \hat{\alpha} + \hat{\beta} * X_i + \hat{u}_i$$

where $\hat{\beta}$ is the estimated effect of X on Y, $\hat{\alpha}$ is the estimated Y intercept, and $\hat{u}_i$ is the estimated error for each observation (also called the residual).

Estimating a bivariate regression model in R uses the `lm` command. The structure of the command is such: `lm(y~x, data=DataName)`. Notice that this formula always has the y (or dependent variable) before the ~ and the x variable (or independent variable) afterward. For the example here, we will be using the American National Election Study subset that we used last time. Recall setting your `working directory` and loading the data:

```
#Setting the working directory
setwd("/Users/sarahhunter/Desktop/Data")

#Loading CSV data
mydata<-read.csv("nes2004subset3.csv")
```

Once you have the data loaded into R, you estimate a linear regression. In this example, the dependent variable is the respondent's rating of President George W. Bush on the Feeling Thermometer scale (where higher values). Our Independent variable will be the respondent's income, making our regression equation:

$$bush\_therm_i = \hat{\alpha} + \hat{\beta} * income_i$$

where $\hat{\beta}$ is the estimated effect of income on the bush_therm (Bush Feeling Thermometer rating), $\hat{\alpha}$ is the estimated Y intercept.

We estimated the model using the following code, remember to name the model in order to create an object (as we did from our first lesson). Also, refer to the section above that explains how the

```
thermometer_mod<-lm(bush_therm~income, data=mydata)
```

Notice that you do not get any output from the above code. This is because you have create and object and assigned it the value of your model. Now that is stored in R. You need to do a couple extra steps to get to the results.

Summarizing and Interpreting Results

To see the results of our model, we can again use the `summary` command.

```r
summary(thermometer_mod)
```

```
##
## Call:
## lm(formula = bush_therm ~ income, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.364 -26.877   5.603  28.991  52.546
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46.9584     2.7734  16.932  < 2e-16 ***
## income        0.4959     0.1721   2.882  0.00403 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.61 on 1064 degrees of freedom
##   (141 observations deleted due to missingness)
## Multiple R-squared:  0.007745,   Adjusted R-squared:  0.006812
## F-statistic: 8.305 on 1 and 1064 DF,  p-value: 0.004033
```

The output of the `lm` command is shown above. The "Estimate" column is the estimated $\hat{\alpha}$ and $\hat{\beta}$ from the regression equation above. This makes 0.4959 the effect of income on the bush_therm score. The more formal interpretation is: "A one unit increase in income leads to a .4959 increase in the respondent's feeling thermometer rating of President George W. Bush".

The second column of that output is the Standard Errors for the α and β estimates. The third column is the result of the t-test for statistical significance. The last column is the p-value. This the the measure of statistical significance most used for regression coefficients. The p-value tells you the probability of find your results (or something more extreme) if the null hypothesis were true. In other words, how likely are your results due to random chance? If the answer is "not likely", we can conclude that this is a systematic relationship. The industry standard is for p-values less than .05 being considered "statistically significant" In our regression, we can see that there is a statistically significant relationship positive between income and Bush's feeling thermometer score.

The other information R gives your is the Residual Standard Error and the number of degrees of freedom (the number of observations used in the model minus the number of parameters estimated). R also provides a measure of fit. The Multiple R-squared and Adjusted R-squared scores tell us how much of the variation in Y is explained by our model. The Adjusted R-squared is a measure that penalizes the model for each parameter added.

Multiple Regression

The Bivariate regression model is simple, however, cannot give any evidence for causality. Political and social phenomena are multi-causal, meaning there are many causes for the same event. Therefore, in order to say that our relationship is not spurious, we must control for confounding factors. We do this by including them in our regression model. In terms of our generic regression equation, we simply add the new variable to the equation:

$$Y_i = \hat{\alpha} + \hat{\beta}_1 * X_i + \hat{\beta}_2 * Z_i + \hat{u}_i$$

where $\hat{\beta}_1$ is the estimated effect of X on Y holding Z constant, $\hat{\beta}_2$ is the estimated effect of Z on Y holding X constant, $\hat{\alpha}$ is the estimated Y intercept, and $\hat{u}_i$ is the estimated error for each observation (also called the residual).

This is called multiple regression. In R, it uses the same `lm` command. The only difference is that we add additional variations using the `+` symbol. In the context of our example, perhaps we think that education also impacts the respondent's rating of President Bush. We would include education in our regression equation by:

$$bush\_therm_i = \hat{\alpha} + \hat{\beta}_1 * income_i + \hat{\beta}_2 * education_i$$

where $\hat{\beta}_1$ is the estimated effect of income on the bush_therm (Bush Feeling Thermometer rating) holding education constant, $\hat{\beta}_2$ is the estimated effect of education on the bush_therm (Bush Feeling Thermometer rating) holding income constant, $\hat{\alpha}$ is the estimated Y intercept.

To estimate a multiple regression in R, use the following code:

```
multiple.regression<-lm(bush_therm~income+education, data=mydata)

summary(multiple.regression)
```

```
##
## Call:
## lm(formula = bush_therm ~ income + education, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.081 -26.437   4.679  28.467  61.255
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  55.3680     3.3172  16.691  < 2e-16 ***
## income        0.8699     0.1895   4.591 4.94e-06 ***
## education    -3.2054     0.7080  -4.527 6.64e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.31 on 1063 degrees of freedom
##   (141 observations deleted due to missingness)
## Multiple R-squared:  0.02652,    Adjusted R-squared:  0.02468
## F-statistic: 14.48 on 2 and 1063 DF,  p-value: 6.261e-07
```

The interpretation of a multiple regression model is very similar to that of a bivariate regression, with one glaring exception. Notice that we described $\hat{\beta}_2$ as the estimated effect of education on the bush_therm (Bush Feeling Thermometer rating) holding income constant. Therefore, we cannot say that a one unit increase in education leads to a 3.2054 decrease in the respondent's score of Bush on the feeling thermometer. This would have been accurate under the bivariate regression context. But, in multiple regression, the estimated effects are only the partial effects. Therefore, to interpret a multiple regression coefficient, we need to say "a one unit increase in education leads to a 3.2054 point decrease in the respondent's score of Bush on the feeling thermometer, holding income constant". The correct interpretation of the income coefficient then is "a one unit increase in income leads to a 0.8699 point increase in the respondent's score of Bush on the feeling thermometer, holding education constant".

Displaying Results with Tables

After you estimate a linear model, R has some cool features to let you convert those results to a professional looking table. You can use a package here called Stargazer (Hlavac 2018). To install the package, use the same code that we have used previously, remembering to use the library command to call up the package. Also, remember to remove the # when running your own code.

```r
#install.packages("stargazer")
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

Stargazer is a package that can create professional looking tables and export them into several different formats. The two you should use are the HTML or text (unless you are familiar with a type setting software called LaTeX). You can get stargazer to export a table for you. You can use the following code:

```r
stargazer(multiple.regression, type="html", out="stargazer_table.doc",
          covariate.labels = c("Income", "Education"))
```

This file will be in your working directory. There are many more ways to customize a stargazer table. This website can help with the options: https://dmyee.files.wordpress.com/2016/03/table_workshop.pdf.

You can also display more than one model in a single table with Stargazer. You can do this by simply adding another model to the stargazer command:

```r
stargazer(multiple.regression, thermometer_mod, type="html",
          out="stargazer_table2.doc",
          covariate.labels = c("Income", "Education"))
```

Another package that makes nice tables in R is `sjPlot`. This is a handy package that does other things such a create predicted probability or marginal effects plot. However, for now, it also makes very nice tables for regression results. The following code shows how to create these tables. You will need to install a few packages for this to work.

```r
#Install all packages (commented out here because I had them)

#install.packages("sjPlot")
#install.packages("sjmisc")
#install.packages("sjlabelled")

#Always remember to load packages as you need them

library(sjPlot)
```

```
## Learn more about sjPlot with 'browseVignettes("sjPlot")'.
```

```r
library(sjmisc)
library(sjlabelled)
```

```
##
## Attaching package: 'sjlabelled'

## The following objects are masked from 'package:sjmisc':
##
##     to_character, to_factor, to_label, to_numeric
```

```r
#To create the table:

tab_model(multiple.regression, thermometer_mod, pred.labels =
            c("Intercept", "Income", "Education"),
          dv.labels = c("Bivariate Regression",
                        "Multiple Regression"),
          string.pred = "Predictors", string.ci = "C.I. (95%)",
          string.p="P-Value", show.ci=FALSE)
```

If you run this code, you will be able to get a table in the right hand side of the R Studio screen, in the Viewer.

Displaying Results with Figures

In addition to regression tables, a good way to show regression results are to use figures. The "effects" package can make a graph of your regression results and includes the confidence intervals. To use this package, first install the package and then load the library into R as usual with a new package.

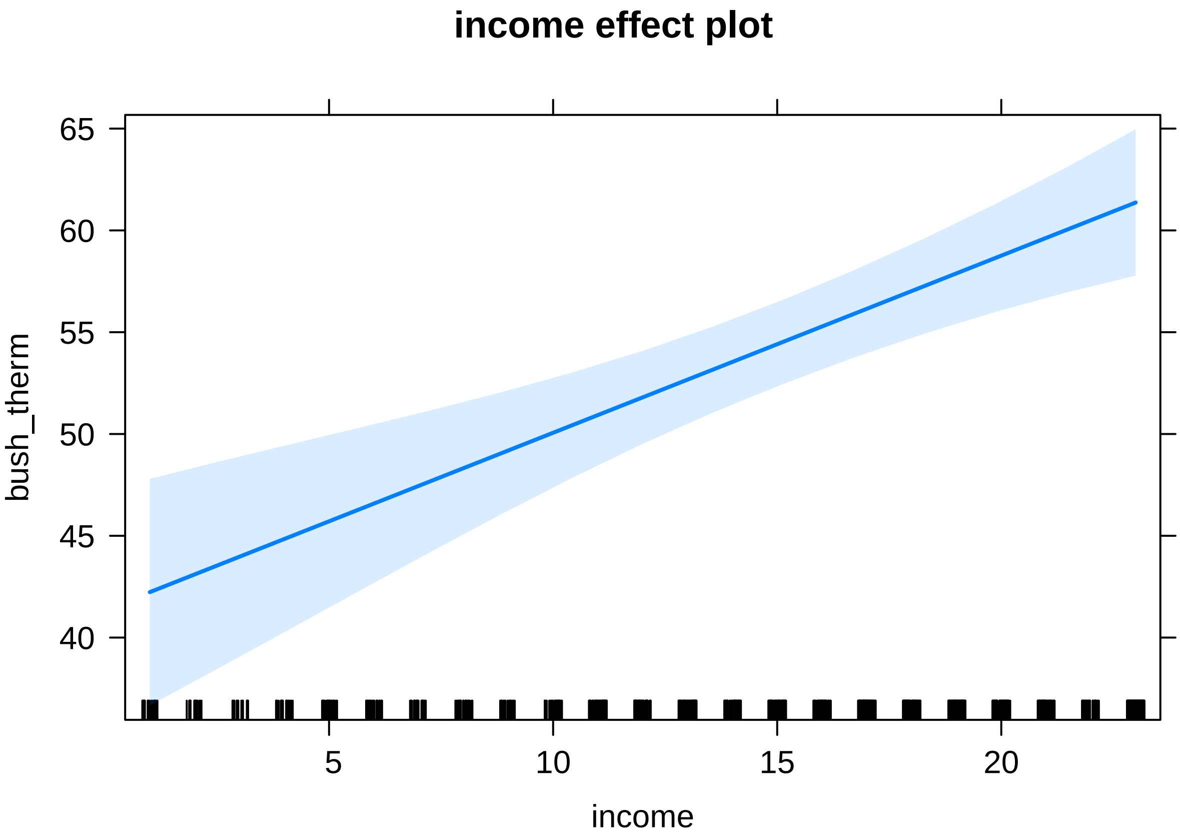
```
#install.packages("effects")

library(effects)
```

```
## Loading required package: carData
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

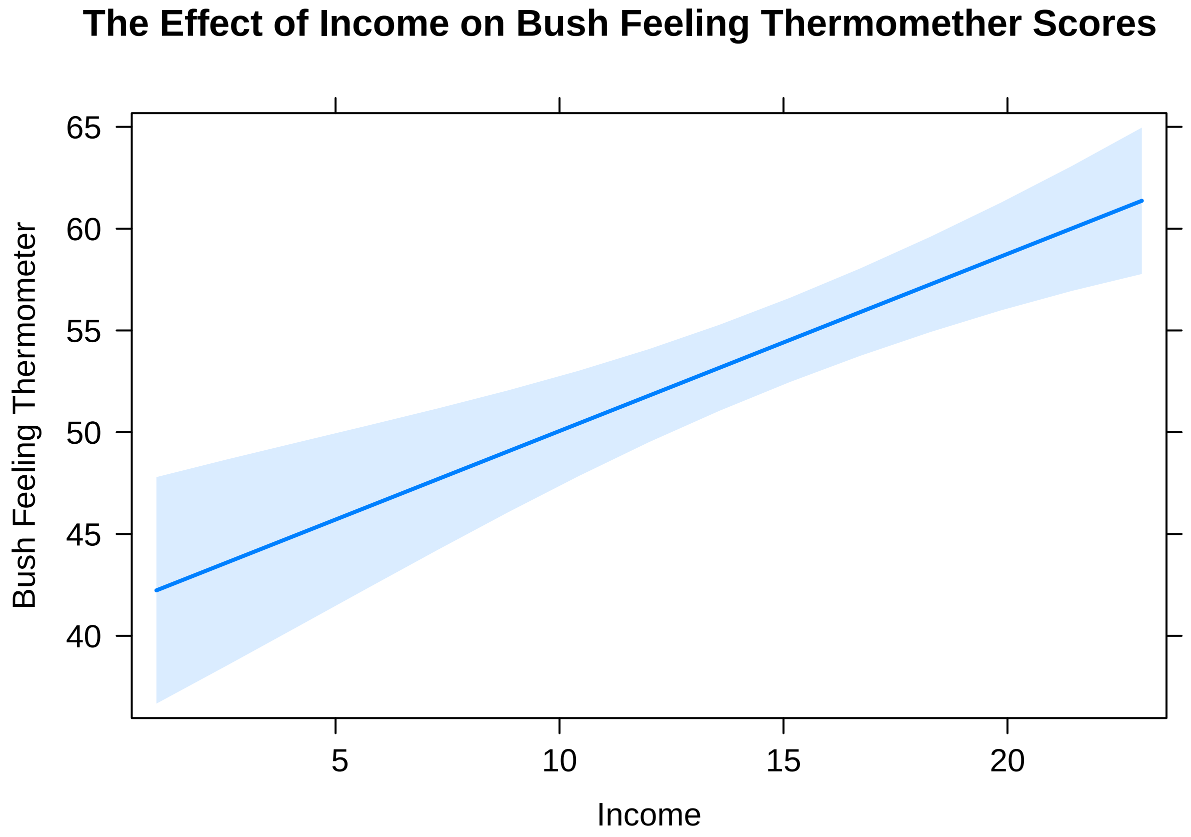
After you load the library, you can then make an effects plot. The basic effects plot is in the following code. This plot gives the the predicted value of the dependent variable (in this case, the Bush Feeling Thermometer score) at levels of the independent variable (income in this plot).

```
plot(Effect("income", mod=multiple.regression))
```



The plot above shows the predicted vale of the Bush Feeling Thermometer score, based on t respondent's income, holding education constant. We can customize this plot with the options, similar to when we covered basic plots. The following code makes the plot look nicer:

```r
plot(Effect("income", mod=multiple.regression), rescale.axis=F,
     ylab="Bush Feeling Thermometer", xlab="Income",
     main="The Effect of Income on Bush Feeling Thermomether Scores", rug=F)
```


The Effect of Income on Bush Feeling Thermomether Scores
Bush Feeling Thermometer
Income